DOCUMENT RESUME

ED 279 683 TM 870 070

AUTHOR Jones, Calvin C.
TITLE Relationships between the National Assessment of Educational Progress and the National Educational Longitudinal Studies Program.
INSTITUTION National Opinion Research Center, Chicago, Ill.
PUB DATE 2 Sep 86
NOTE 74p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094. Document contains blurred type. Best copy available.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Agency Cooperation; Comparative Analysis; Computer Assisted Testing; *Data Collection; *Educational Assessment; Elementary Secondary Education; *Information Utilization; *Longitudinal Studies; Measurement Objectives; *National Surveys; Program Costs; Research Design; Sampling; Testing Problems; Testing Programs
IDENTIFIERS High School and Beyond (NCES); *National Assessment of Educational Progress; National Longitudinal Study High School Class 1972

ABSTRACT

The National Assessment of Educational Progress (NAEP) is compared to the two studies of the National Educational Longitudinal Studies (NELS) Program--the National Longitudinal Study of the Class of 1972 (resurveyed in 1973, 1974, 1976, and 1979) and the 1980 High School and Beyond study (partially resurveyed and retested in 1982). The third phase of the NELS program, the National Education Longitudinal Study of 1988 (initiated in 1986) breaks new ground. Transitions into and through secondary school and the processes of program selection and tracking will be studied, and a less biased sample of dropouts will be included. There are advantages and disadvantages to greater integration and exchange of information between these studies. Five factors influencing the utility of the data collected in these two programs are reviewed: (1) Population definitions and sample designs are inconsistent and change over time. (2) The response burden upon institutions and individuals can be lessened by NAEP's use of matrix sampling and BIB spiralling. Coordination is needed to prevent schools from participating in both surveys at once, and to help in data exchange. (3) Data should be useful to educators and administrators at the state and local level. NELS allows states to survey supplemental data; this may also be useful for NAEP. (4) Data collection and processing costs may be helped by states' supplementary data collection and by increased efficiency. Otherwise, the required sample sizes would be too large and too costly. (5) Opportunities provided by new technology involve microcomputers, computer assisted testing, and computer assisted interviewing. (GDC)

ED279683

Relationships Between the National Assessment of Educational Progress and the National Educational Longitudinal Studies Program

Calvin C. Jones

NORC

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. C. Jones

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper commissioned by

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

TM 870 070

BEST COPY AVAILABLE

RELATIONSHIPS BETWEEN
THE NATIONAL ASSESSMENT
OF EDUCATIONAL PROGRESS

AND

THE NATIONAL EDUCATIONAL
LONGITUDINAL STUDIES PROGRAM

NORC
UNIVERSITY OF CHICAGO

RELATIONSHIPS BETWEEN

THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

AND

THE NATIONAL EDUCATIONAL LONGITUDINAL STUDIES PROGRAM

Calvin C. Jones
Senior Survey Director

NORC

September 2, 1986

Relationships between the National Assessment of Educational Progress and the National Educational Longitudinal Studies Program

OVERVIEW

Since its inception in the late 1960's, the National Assessment of Educational Progress (NAEP) has had as its primary objective the measurement of learning across a broad range of academic subjects and skills by elementary and secondary school-aged children. Although the core of the research design is the collection of standardized test data from large, nationally representative samples of students, the research objectives of the study and consequently its overall design have evolved considerably over the years. NAEP now invests substantial resources in a data collection design that will permit the study of relationships between assessment scores for different subjects or exercises. In addition, NAEP now routinely collects data on students' backgrounds and schools, as well as information from samples of students' teachers and parents for use in explanatory analyses of achievement scores. Oversamples and supplementary samples have been included to support research on students at lower socioeconomic levels, language minorities and racial and ethnic subpopulations, and young adults not enrolled in schools. Collection of sampled students' school records (transcripts) is now under consideration. The products of NAEP have been extended far beyond the issuance of periodic "national report cards." The kind and complexity of analyses possible with NAEP data will probably continue to expand in response to

pressing demands for information on the impact of specific educational policies and practices.

During much of this same period, the National Education Longitudinal Studies (NELS) program has undergone an even more dramatic sequence of design changes as the information needs of education policy makers have grown. Beginning with the National Longitudinal Study of the High School Class of 1972 (NLS-72), a survey of over 22,000 high school seniors of that year, the program's main focus was the study of postsecondary school access and choice, and the impact of these educational activities on careers and other aspects of individual development. The Base Year survey included collection of detailed questionnaire data as well as scores on a battery of cognitive tests. After the Base Year, the full NLS-72 cohort sample was resurveyed four more times (in 1973, 1974, 1976 and 1979), and a substantial subsample of the original cohort (15,000 cases) is being resurveyed in 1986.

In 1980, the second major NELS study, High School and Beyond (HS&B), was launched. With its large sample of 1980 twelfth graders, HS&B sought in part to replicate the design and update the findings of NLS-72 to the decade of the 1980's. It also expanded its scope to the study of growth through the last years of secondary school by including a sample of 1980 tenth graders. Like NLS-72, the Base Year survey of HS&B collected questionnaire and test score information from large samples drawn from over 1,000 public and private secondary schools. The younger cohort was resurveyed and retested in 1982 (when most were twelfth graders). In

addition, the inclusion of the younger cohort permitted the study of a substantial number of high school dropouts, who were retained in the sample for all followup surveys. Since leaving high school, subsamples of both cohorts have been resurveyed at two year intervals.

Besides student surveys, HS&B also collected detailed data on the students' school environments, obtained secondary school and postsecondary school transcripts and postsecondary financial aid records for linking with questionnaire and test data, and collected a limited amount of data from the parents of subsamples of students in each cohort.

The Base Year of HS&B also established the process of extending to state education agencies and other Federal agencies the invitation to supplement (at very low marginal cost) the core sample and survey design to provide for state level or other subpopulation representativeness in the database. The data resource sharing process was later broadened to include all public and private research agencies, who were permitted to supplement the studies by sponsoring special samples or adding questionnaire items at low cost, provided they accepted all Department of Education standards concerning privacy and confidentiality of responses and timely public release of datatapes. This cooperative approach greatly enhanced the value of the HS&B datafiles and provided the basis for a broad funding consortium to support the Fifth Followup survey of NLS-72.

In early 1986, the third study of the NELS program, the National Education Longitudinal Study of 1988 (NELS:88) was initiated. Although

NELS:88 will be based squarely on the methodology of NLS-72 and HS&B, it also breaks much new ground. First, by beginning with a cohort of eighth graders in 1988, the design will permit the study of transitions into and through secondary school, and the processes of program selection and tracking now believed to be crucial elements of that transition. Second, starting with an eighth grade cohort will ensure that the initial sample will include a more complete, unbiased sample of individuals who will eventually drop out of school.

In terms of cumulative dollars expended and amount of data collected, NAEP and NELS are two of the largest research efforts sponsored by the U.S. Department of Education. Each of the studies was established to advance a specific set of research goals that required unique design features, and these basic differences in purpose persist to the present day. However, at a very general level, the two research programs also appear to share many common objectives and methods. Both use sample survey methodology on a large scale to collect nationally (and subnationally) representative, policy-relevant data from institutions and individuals on educational processes and outcomes in elementary and secondary schools. Both collect data on environment and social and economic background to analyze students' achievement levels, measured by performance on standardized cognitive test batteries. Both are concerned with assessing change in educational accomplishment (test scores) over time, although each in a different way. Finally, both programs have produced substantial and growing bodies of literature, volumes of statistical estimates and substantive findings, and have attracted the

attention of numerous highly skilled data analysts, who are prepared to exploit each new wave of data.

Yet, despite their common origin in the legislative mandate of the former National Center for Education Statistics (now the Center for Statistics of the Office of Educational Research and Improvement), for several years the two programs have operated virtually independently. Until recently, the studies were monitored by separate agencies in the Education Department, were conducted by different contractors or grantees, and benefitted only from the minimal integration that resulted from a few common members of their respective non-governmental advisory bodies. They have gathered data from independent (and uncoordinated) samples of schools, administrators, and students, and have developed research philosophies and measurement techniques that capitalize upon each study's unique design and emphasize its particular objectives.

The result is that while each of the two studies has extraordinary power to address specific types of research or policy questions, ultimately the findings from each add little to policy makers' ability to use information from the other. If anything, over the past sixteen years the extent of the design articulation between NAEP and NELS has decreased as the two studies have become more specialized. Recently, leadership at OERI and CS has called for greater integration among all CS studies. This has led to significant expansion in the consultation and coordination between the firms responsible for questionnaire and cognitive test development for NELS:88 and the next NAEP assessment. However, given the

current objectives of the two programs, increased emphasis on the sharing of questionnaire and test items is unlikely to produce dramatic changes.

My purpose in this paper is to call attention to several design, logistical, data utilization and policy issues that affect the operation of both the NAEP and NELS programs, and to outline the advantages and disadvantages of greater integration between these important studies. I have limited consideration to the following issues that I believe have the greatest impact upon the overall utility of the data collected by the two programs:

1. Population definitions and sample designs;
2. Response burden upon institutions and individuals;
3. The usefulness of data and findings to educators and administrators at the state and local level;
4. Data collection and processing costs;

and 5. Opportunities provided by new technology.

Each of these topics relates to several of the others. Clearly, certain alternative designs for NAEP could dramatically affect the utility of its data to state education agencies, but also could greatly increase the costs of the study. Similarly, the application of newer data collection and dissemination technologies could reduce respondent burden, expand access to findings, and possibly lower certain costs. However, the most important thread linking these topics is the assumption that neither NAEP nor NELS is merely a basic statistical indicators program. All are facets

of the larger issue of how large scale Federal research programs should be organized to provide the responsible actors with the tools to evaluate, change, and improve educational programs and policies.

1. POPULATION DEFINITIONS AND SAMPLE DESIGNS

Student Populations

The primary subjects of both NAEP and NELS are elementary and secondary school students. Given the different but potentially complementary emphases of the two projects, it seems desirable for both studies to collect their specific types of data on the same student populations. Unfortunately, each study defines its populations somewhat differently, limiting the use of findings from one to further understanding of relationships observed in the other. NAEP, partly in order to avoid the appearance of attempting to establish national grade norms against which local systems might be compared, has (up until the most recent assessments) defined its populations as single-year age cohorts (9, 13, and 17 year olds). Since its beginning, the NELS program has defined its populations as grade cohorts without regard for age.

The most recent redesign of NAEP has introduced dual sampling (both by age and by grade) that provides the opportunity for significantly increased comparability with NELS studies. However, the design also introduced changes in the reference dates for defining the two younger age cohorts, so that the modal grades for all three age cohorts in the next assessment (grades 3, 7, and 11) are inconsistent with those sampled by NELS:88 (8, 10, and 12).

For the 1983-84 assessment, the modal grades for 9 and 13 year olds were grades 4 and 8. This relationship holds if the definition of an

age-eligible student includes those who attain the specified ages during the January 1 to December 30 calendar years as in earlier assessments, but is false under the newly adopted reference dates of October 1 to September 30. Compared to the old specification, the new design selects younger students for the age cohorts. Consequently, the modal grades for 9 and 13 year olds in later assessments are reportedly grades 3 and 7, rather than 4 and 8.

NELS:88 will draw a nationally representative sample of eighth graders in the 1987-88 school year. There will be very little overlap with the 13 year old/7th grade population sampled for the NAEP assessment of that year. A small but significant proportion of the NAEP samples will still be drawn strictly on the basis of age, so that even under the new reference dates, some eighth graders will be selected. In general, however, it does not appear that the defined populations for NELS:88 and future NAEP assessments will coincide, except in an excessively restrictive sense.

Figure 1 illustrates the problem of the inconsistent designs, focusing on the NELS:88 grade cohort and the NAEP 13 year/7th grade cohort. The eighth grade cohort sample for NELS:88 may include students that span up to six years of age, including students who are one to two years below the modal age for a grade (because of early school starts and grade skipping), as well as students who are up to three years above the modal age (due to late starts, retention in grade, lower grade placement of limited-English proficient immigrants, etc.). By choosing a

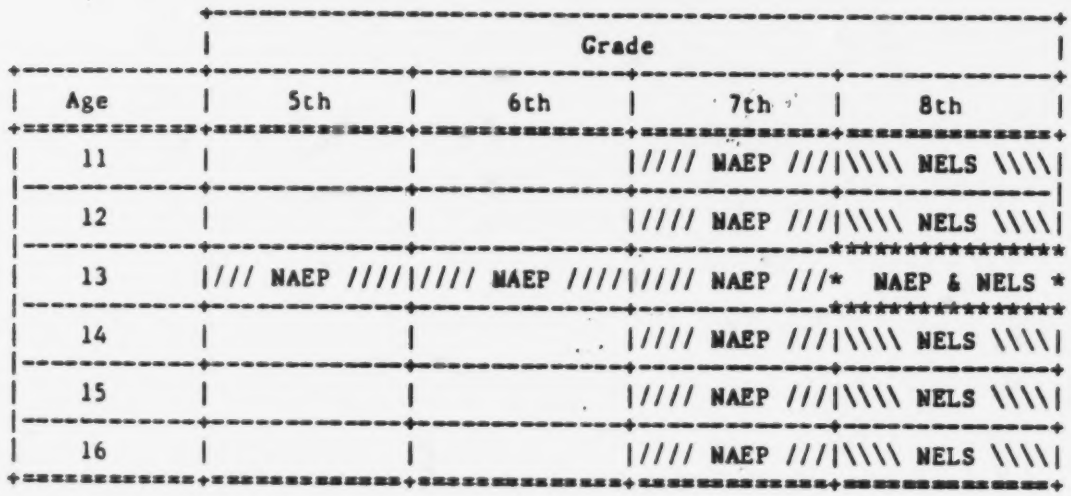
significant part of its sample from single-year age groups, NAEP samples may include students who are one grade ahead or up to two grades behind the modal grade for the age cohort. Since the defined modal grade for NAEP 13 year olds is grade 7, compared to NELS:88's grade 8, and the samples for the two studies will be drawn in the same academic years, the overlap in the population definitions will be limited to that portion of NAEP's 13 year old cohort who are beyond the "expected" grade for normal progression students. While it is difficult to estimate the actual numbers without more information, projections for the 1983-84 assessment indicated that about 20 percent of each age/grade cohort would be selected for age only (i.e., not in modal grade). Since some portion of these 13 year olds will be in the sixth or lower grade, it is reasonable to estimate that only 10 to 15 percent of the NAEP 13 year old sample will belong to the same grade population as the NELS:88 eighth grade cohort. The extent of the overlap would be much higher if the reference dates were changed so that the population included a younger birth cohort, but would only be complete if the NAEP design were altered to sample grade cohorts consistent with NELS population definitions.

Figure 1

Relationship between NAEP and NELS Populations

of Age and Grade Level Student Cohorts

(NAEP 13 Year Old/Seventh Grade Cohort; NELS:88 Eighth Grade Cohort)



The arguments in defense of retaining age cohort sampling (primarily comparability with past assessments, the "common meaning" of age across geographic regions, and the problems of interpretability and comparability of grade-average scores for students who are over age for grade placement) seem to me much weaker than those supporting its elimination in favor of grade cohort designs. In fact, the recent emphasis at OERI on integrating its major studies is doomed to a degree of frustration by the continuation of this basic design difference. Increasingly, OERI-CS studies have stressed the incorporation of design features that would enhance the

usefulness of findings to educational practitioners and policy makers at all governmental levels. At minimum, this should require that major studies focus on the administrative and instructional units that are most subject to manipulation: schools, grade-levels and classrooms. Aside from the trends in test scores for age groups now possible with NAEP data, and the generally accepted legal minimum age of 16 years for dropping out of school, very little else in the operation of the nation's educational system -- not school organization, not curriculum development, not ability grouping, not program tracking, not even first entry into public schools -- is determined by the age of the student.

Since the middle of the nineteenth century, the grouping of students by grades has served as the basic structural unit for focusing instruction and assessing progress in elementary and secondary schools. This fact should be fully reflected in the design of the major studies sponsored by the Education Department. The political and social climate is greatly changed since NAEP's founding. Design features should no longer be determined largely by speculation about adverse reaction from state or local officials but should be organized largely to serve information needs and requirements at those levels. During their developmental history, NELS and many other studies have evolved around this dictum. NELS:88, the most recent member of the program, has retained the grade-level sample design of prior NELS studies, and has carried the goal further by following student transitions from grade 8 through 12, expanding the sample as necessary to ensure the representativeness of grade cohorts at each level. The NELS:88 design, like High School and Beyond before it,

has gained substantial enthusiasm, rather than concerned criticism, from state and district education officials. The response to giving NAEP a grade-level orientation would likely be no different.

Under the current situation, NELS:88 has experienced some difficulty in furthering the goals of OERI and CS to better integrate their studies. During the earliest planning meetings for NELS:88, strong emphasis was placed on the goal of ensuring that NELS data collection instruments would be compatible with NAEP (and other surveys) by sharing test and questionnaire items. However, given the differences in populations for which test items are developed for the two studies, both the rationale for the item sharing, and the means for identifying precisely which items should be included in both studies quickly became problematic. Within a relatively short time allotted for testing, the paramount goals of NELS:88 are to measure individual student change and growth from the eighth, to the tenth, and the twelfth grades, to design a test battery with maximal sensitivity to curriculum taught in those grades, and to collect the full complement of test data in four subject matter areas on all sampled students, without matrix sampling, incomplete block designs or other devices. The additional goal of overlapping items with tests designed for the more diverse populations of all 13 year olds and all 17 year olds could be met with only a restricted set of choices. Integration of this type would be more feasible if NAEP tests measured eighth and twelfth grade-level, rather than the age-grade combinations of the current design.

Non-student Populations

NAEP has often been described as a program for producing descriptive, rather than analytical results. In the current grant cycle, the amount of data collected to support limited analyses of test scores has increased. Some of the additional data is collected from participating students, some is obtained from school officials, teachers and (in special cases) from parents. Furthermore, NAEP's reporting requirements have been substantially expanded to include several monographs on crucial policy issues.

Since the beginning, NELS contract requirements have reflected the primacy of the analytical goals of researchers and policy makers at multiple levels. Across the many waves of longitudinal data collection and analysis to date, one design theme has recurred more often than all others: effective longitudinal or cross-sectional analyses of student outcomes requires more complex data structures than can be obtained from student self-reports alone, and that, at minimum, the analysis of educational performance and growth requires data on both the institutional context and home environment. The current NELS:88 design reflects these points to a greater degree than its predecessors. The data needed to estimate complex multivariate models of student status and change have been identified by the analysts using data from the prior studies, and are designed into the Base Year survey.

If the next NAEP cycle is to continue the evolution toward greater utilization for policy formation and evaluation, its designers should

consider further expansion in the kinds and amounts of non-test data collected, even if this implies increasing respondent burden. Very little analysis relevant to educational policy is possible with such small amounts of questionnaire data collected from students on their backgrounds and individual characteristics. Assessing the impact of programs and policies on achievement requires accurate, timely and individual level measures of critical environmental variables. The number of possible analyses of NAEP would be greatly increased by having direct measures of family SES, the features of the home educational support system, school characteristics and policies, classroom and teacher practices, and other background measures on all sampled students, even if the blocks of assessment exercises are distributed by BIB spiralling or other methods. Data of this kind would foster the use of NAEP data for educational improvement, dramatically raising the return on society's investment in the program.

The Out-of-School Population

Because of funding limitations, NAEP assessments have sometimes, but not consistently, included samples of young adults well beyond school age for the purpose of testing functional literacy. Currently, one of the most rapidly expanding interests within NAEP is the assessment of literacy levels in the young adult population, especially youth of high school age (17 year olds) who are not currently enrolled in any secondary school program. Renewed consideration is now being given to procedures for drawing supplementary samples of young adults through the use of area

sampling frames (as was attempted for the 1983-84 assessment) and other approaches.

NELS studies have addressed similar issues, but in a longitudinal context. Rather than selection cross-sectional samples of non-students, High School and Beyond and now NELS:88 have designed in procedures for the longitudinal study of school dropouts. Since the First Followup of HS&B in 1982, the study has included a nationally representative sample of high school dropouts. HS&B tenth graders of 1980 who were found to have left school by 1982 were located, resurveyed and retested using a minor variation of the group survey administration procedures used for currently enrolled students. The response rate for the dropout sample was reasonably high (89 percent of the total), and the questionnaire and test data collected on these youth was completely comparable with that obtained for school completers.

The major difficulty with the HS&B sample of school leavers is that a large fraction of dropouts had already left school prior to sampling activities for the tenth grade cohort in 1980. These early dropouts are known to be disproportionately male, poor, non-white and of Hispanic origin or descent. This problem is one of the strongest justifications for the change in design from the HS&B tenth grade cohort to an eighth grade cohort for NELS:88, which is expected to include a full comparison group of school leavers from the original eighth grade cohort. Whether the NELS:88 sample members who leave school prior to high school graduation might serve as an adequate out-of-school sample for NAEP will

depend greatly upon whether age-level or consistent grade-level cohorts are the population of interest for that research.

Sample Size, Item Sampling and Design Complexity

Related to the need for background data from non-student populations for analytical purposes both NAEP and NELS must confront the issue of the quantity of data and the specific data elements collected from each sampled student, school, parent and teacher. Over its history, the NAEP design has evolved an increasingly complex design for determining the test items and other measurements to be made on each sampled student. Matrix sampling and now BIB spiralling have become accepted aspects of the NAEP design, justified on the grounds that the analysis plans do not require complete data at the individual level to serve the "national report card" function. This justification may still be reasonable as a means for distributing assessment exercises across students and limiting individual response burden. But is there adequate justification for spiralling background questionnaire items across students as well?

Given the original objectives of NAEP, matrix sampling and BIB spiralling have tended to increase the efficiency of the data collection task. However, they have also created two well-recognized problems for NAEP's usefulness to the education research and policy community: 1) the basic comprehensibility of NAEP measurement procedures, and the resulting restrictions on use of NAEP data by researchers, state and local educational officials and practitioners, and 2) real and substantial

increases in the costs of implementing the complex survey procedures, and in the training and facilities required of policy analysts to extract and publish useful information from the NAEP database.

Besides NAEP, few other major studies rely on these kinds of efficiency enhancing devices to such an extent. For example, in contrast to NAEP, the NELS designs for any given base year survey, while by no means simple, have tended to become more straightforward for each subsequent study within the program. The longitudinal analysis goals of the NELS designs and the need to reduce sample sizes in the post-high school rounds of the study increase the importance of having equivalent data for all initally sampled cases. More importantly, however, we have repeatedly found that the simpler designs are more understandable to both administrators and analysts, and, aside from the complexities of the disproportionate sample design and the need for adequate computing facilities to handle large datafiles, the datasets have become more and more "user-friendly." To date nearly 800 books, journal articles, US Department of Education reports, dissertations, and university publications have been produced using NELS datasets

The major features distinguishing the data collection strategies of NAEP and NELS are few but important. To try to avoid the two problems listed above, NELS has moved away from complex data collection protocols, and especially has avoided using samples of questionnaire or test items from the population of possible items used for the entire survey. A single, fixed questionnaire is used for all sampled cases from each

population included in the study. Within a questionnaire, behaviorally different groups (e.g., employed vs. not employed; postsecondary school attenders vs. non-attenders) are handled by routing or "skip" patterns in the flow of items. In the final datafiles, all observations are intended to have equivalent measures on all survey or test items or composite variables. Furthermore, if data are to be collected from populations linked to students (e.g., parents or teachers), every attempt is made to collect equivalent linked data for every sampled student, rather than for subsets.

This strategy is feasible when the total number of desired items per case can be controlled and negotiated to stay reasonably close to the budget constraints of the contract. It would be impossible to maintain if the study was somehow forced to measure six times the number of survey items in the same amount of respondent burden. While NAEP may be obligated to include very large numbers of test items, it is doubtful that the number of background data items required for analyzing assessment scores would present the same dilemma.

The alternatives of matrix sampling, incomplete block designs, and restricting collection of specific data elements to subsamples are often justified as being more cost-effective or efficient means to collect more different kinds of data for a fixed cost. However, complicating the data collection design is not a panacea. These techniques have been most effectively used in "one-shot" studies with narrowly defined objectives and limited resources. BIB spiralling may be the wisest choice for

distributing the assessment exercises. For large scale, repeating, multi-purpose studies, hidden costs eventually emerge from behind these techniques.

First, although they seem to have been implemented with reasonable success to date, these techniques are more difficult both to design and to carry out in the schools and classrooms. Developing the BIB design imposes higher front-end costs. Printing and assembling relatively small quantities of a large number of different booklets raises costs considerably. Compared to simpler designs, training, supervision, and quality control costs must be higher when these devices are used. In addition, the added complexity raises the possibility of loss of control over the administration of the survey, or of survey cost overruns incurred in preventing such problems. Furthermore, the processing of complex data structures into analysis files and public-use data tapes is made more difficult and hence costly. The opportunities for processing errors, and the delays and additional costs of correcting them, are significantly increased.

Second, the data from matrix samples and BIB designs are not accessible to many researchers, who may lack the training, staff, computing facilities, or interest in the methods for coping with very complex data structures. I believe that design features that limit the accessibility of research resources to the widest possible spectrum of analysts represent heavy opportunity costs to society by inhibiting the kind and number of contending analyses that may contribute to our

understanding of educational processes.

Third, even for those who are capable of exploiting the data, the additional setup and computing costs of handling complex weighting schemes, data relationships, imputations and other missing data handling techniques, and estimation procedures can sometimes offset the costs that might have been incurred collecting identical data from all sampled cases. High costs and complexity of analytical procedures inhibits replicability, simplicity and clarity of presentation, and rapid accumulation of useful findings.

Finally, since under these item-sampling schemes many data elements are collected for discrete (though probabalistically determined) minorities of sampled students, it is often impossible to analyze relationships within important ascriptive or behavioral subgroups because of the small numbers of cases with necessary data. The hidden cost of this limitation on the generalizability of findings to specific subpopulations of students or schools is impossible to estimate.

Given the very real limits on resources for the collection of assessment scores and necessary background data for its interpretation, NAEP sponsors must place a high value on design efficiency. OERI must favor designs that produce the widest variety of high priority data for a given budget. However, in evaluating data collection strategies, greater consideration should be given to the impact of alternative designs on the costs of research and data utilization activities -- many of which are not

included in the initial project budget. A brilliant design that generates a data structure too complex and costly for widespread exploitation may fill computer tape libraries, but may fail to fill scientific, professional, and administrative journals and public information releases with useable information on how teaching and learning may be improved.

2. RESPONSE BURDEN UPON INSTITUTIONS AND INDIVIDUALS

Institutional Burden

Both NAEP and NELS are large studies conducted at regular intervals. In fact, under the current schedule, both studies will have major field efforts underway in school systems in the 1987-88 academic year. The NELS:88 sample design currently includes 1,000 public and private schools containing eighth grades. Although the NAEP sampling plan for the next round is not yet final, survey operations are expected in approximately 1,500 schools, divided among elementary, middle (or junior high) and secondary (senior high) schools. Although the real probability of any individual school or student being selected into both the NAEP and NELS samples is very small, the overlap between the study schedules tends to create a perceived response burden in the minds of education officials at the state and local level that should not be ignored. The perceptions of these powerful gatekeepers concerning the costs and benefits of disruptions to the instructional schedules of the schools they administer can profoundly affect the success of the survey.

Simply by virtue of the sizes of the studies and the way the population of students is geographically distributed, both NAEP and NELS will conduct operations in all or nearly all states. Chief state school officers will be approached by the staffs of the NELS contractor and NAEP state agencies to obtain support in gaining access to districts and schools. It is most essential that the state level contacts for these studies be

coordinated by the data collection firms that are selected.

Both studies use multi-stage sample designs that include schools (and, by implication, school districts) as intermediate stage units to select samples of students. Given the large number of school buildings, if the two studies were to draw simple random samples of schools, the probability that any school would be selected for both studies would be close to zero. However, since both studies want student samples that are as close to self-weighting as possible, schools are selected with probabilities proportional to their sizes (enrollments). This constraint substantially increases the probability that the largest public schools containing seventh and eighth grades in the largest districts in the most populous states will fall into both samples. School sample selection for NAEP and NELS should therefore be coordinated as a Federal requirement. Otherwise, staffs for the organizations conducting both studies may wind up competing for access to sampled districts and schools to the detriment of both studies.

Individual Respondent Burden

Because of the numerous sensitive aspects surrounding the initiation of a national testing program, extraordinary measures were invoked to minimize the response burden on individual NAEP sample members and upon the schools they attend. The project description suggests that student

level data collection is limited to what can be accomplished within a single standard length class period (45 to 50 minutes). Data on student backgrounds are also assumed to be collected during this brief interval. Although the student level burden for the NELS studies has been significantly longer on average, it too has been limited to approximately one hour of testing (for four or more subject areas), and an additional hour for the collection of questionnaire data.

Compared to other Federally sponsored survey research programs, this level of respondent burden is close to the overall average for all types of studies. However, in light of the declared goals of the two studies, and the high value currently placed on precise, detailed information on the educational system, this level of response burden is restrictively small. It is paradoxical that so much effort should be spent limiting response burden when, logically, the need for education data at all levels of government suggests that individual burden levels should be modestly increased. Moreover, given the large percentage of total costs that are fixed (related to the development of the tests and questionnaires, and related to gaining access to and holding survey administrations at the sampled schools), the marginal costs of expanding the amount of data collected from individuals is an efficient way of improving the study.

Increasing response burden may risk lowering participation rates by sampled individuals. However, there are several mitigating circumstances in the NAEP and NELS designs. First, considerations of burden are likely to be more central to the support and participation decisions of state

school officials, district superintendents and school principals than to those of individual students. If perceived burden upon students is higher, these officials might show less willingness to permit survey activities in their states, districts and schools, unless the increase in burden is offset by perceived benefits. Exactly what type of perceived benefit will work is difficult to predict. However, throughout the history of the NELS program, the single most frequently cited benefit requested by officials at all three levels has been information about the organization or geographic area for which they are responsible. I return to this point in the following section.

As suggested above, matrix sampling and BIB spiralling were introduced into NAEP for the purpose of enlarging the research potential of the study while limiting the response burden on individuals. While these methods increase the number of types of estimates that can be made from a sample, they also impose costs on the ability to analyze and generalize from the data. Now that some experience has been gained with the use of these techniques in the 1983-84 assessment, their overall benefit to the program should be carefully evaluated. In particular, the complexity of the data collection design should be considered in light of the alternative of permitting significant increases in individual response burden. The NELS program has generally succeeded in its data gathering objectives while imposing about twice the response burden that NAEP now places upon individual students. NAEP designers should reconsider the tradeoffs between design complexity and the value of obtaining needed background data from all sampled students.

Response Burden as Duplication of Effort

One of the most frequent objections made by state and district officials to Federal data gathering efforts like NAEP and NELS is that they appear to duplicate the collection and processing of information that is already being gathered in a similar form by state or local administrative or research procedures. Some of these information systems are well designed and sophisticated, and serve local information needs as well as supply required information to Federal program monitors. "Why," it is asked, "is it necessary to collect all of this again in a slightly different form?"

Given the assumptions behind the question, few of the usual justifications are very satisfying to those who pose it. Whether or not the differences in purpose of the local and Federal research data systems are understood, questions of this kind are likely to be more frequently and militantly put to large studies like NAEP and NELS in the future. In part, such questions also ask "Why is the Federal government collecting so much test data and other information on student performance when it has no responsibility in this area? Why can't they make use of the information systems and research efforts already in place?"

Since NAEP and NELS are voluntary sample surveys, this problem, even more than that of individual response burden level described above, can only be addressed successfully by close coordination with state and local officials during project development, on the one hand, and the introduction of meaningful incentives such as the reporting of useable

small area data from the studies to these officials, on the other. The current design of both NAEP and NELS precludes feedback of this sort in all but a few instances of state representative NELS samples. The following sections some possibilities for greater information sharing.

3. THE USEFULNESS OF DATA AND FINDINGS TO EDUCATORS AND ADMINISTRATORS AT THE STATE AND LOCAL LEVEL

There are few issues more sensitive than that of whether and how state level estimates of educational progress can be obtained by NAEP that will be valid, scientifically comparable, politically acceptable, and affordable. Given the role of the states in developing and implementing educational practices, there is little question that comparable state-level data is needed. However, the problems of how it should be generated have been repeatedly identified and will not be restated here. My purpose is limited to describing the procedures used by NELS studies to begin to build serviceable bridges between the Federal "core" studies and fully comparable supplementary studies sponsored by state research departments. The successes, partial successes and failures of that enterprise may be useful to NAEP designers.

The approach to state supplementation of NELS studies is to offer states the opportunity to augment either the sample sizes or survey questionnaires (or both) in order to provide state researchers with data to accomplish their specific goals. Prior to each study, the sponsoring Branch within CS sends to each state office a projection of the sample sizes and estimates of the statistical power that each state would have within the "core" sample, and establishes rules for supplementation that will permit a state to be considered a separate sampling stratum. Estimated costs are discussed as well. Because the developmental effort to design the procedures, questionnaires, and tests is covered by the

contract for the national study, the states need only bear the marginal costs of adding data items for each sampled student, or of adding schools and students in order to obtain more precise statistical estimates and/or desired subgroup comparisons. Because such a large portion of the basic costs are covered, many states find that the costs of supplementing the core to provide state representativeness is between 25 percent and 33 percent of the cost of a stand-alone state study of comparable size. Beginning with High School and Beyond, only Illinois, Washington and Ohio chose to sponsor state supplements. To date, however, over a dozen states (or city districts) have expressed interest in supplementing NELS:88.

There are many reasons to imagine that the supplementation rate for NAEP should be higher than for NELS studies. Perhaps the most germane is that the NELS studies are designed to produce their greatest payoffs several years after the Base Year survey is launched. The time frame is probably too long to attract much attention from officials facing many immediate problems. Useful NAEP information is produced on a much shorter time schedule and should generate more interest. In at least one case, however, NELS:88's longer time frame was itself the principal attraction for a state education department. In this state, the NELS:88 cohort will be the first scheduled to graduate from high school (in 1992) under a comprehensive set of policy initiatives recently enacted. The NELS:88 design fits the state's interests and schedule perfectly; it has chosen to sponsor a large supplement in order to pursue its special needs, despite its already large Federally-sponsored core sample.

If state supplementation is to become much more common, I believe that the level of interaction between the state and Federal offices (or the study contractor) needs to be more intensive than has been possible to date. At present, state officials are given basic information, and left to decide whether to take further steps. More information, advice and planning assistance should be provided to state offices in order to clarify the benefits of supplements. State researchers will only rarely have the time or staff to pursue the information on their own. Even the ideal fit between the NELS:88 design and the research interests of the state described above might have gone unnoticed if the state's research director had not been a member of the NELS:88 Advisory Panel.

In contrast, the focus of NAEP has potentially much greater immediate appeal to state officials. And, increasingly, state research departments are becoming less wary of the idea of state level comparisons of assessment scores, and more willing to collaborate in efforts to make them useful and appropriate. The establishment of the Committee on Assessment and Evaluation of the Council of Chief State School Officers, with its mandate to explore just these issues, indicates just how much attitudes have changed since the late 1960's.

NAEP designers should consider the inclusion of a requirement in the next NAEP cycle for the grantee and the Federal monitoring office to make much greater efforts to work with state assessment officials in order to attract as many state supplements as possible to the NAEP core. Many prior recommendations have discussed the provision of "technical

assistance" to state assessment groups, vaguely hinting that states were less capable of mounting local efforts of suitable quality. In fact, much of the exchange necessary to building state-Federal partnerships is likely to be technical assistance by state officials to Federal program managers in their efforts to understand sufficiently the relationships between state curriculum programs and the contents of their assessments. For this to occur, the outreach effort by the NAEP staff must be increased.

The ability to attract state supplements has the potential for solving several of the pressing problems discussed above. As suggested in the following section, state cost sharing is the best hope for surmounting the fiscal obstacles to providing state level data. In addition, as has been true for NELS surveys, state participation in a Federal study can have beneficial impacts on the conduct of the core survey. Implementation of the data collection task is made more effective and cost-efficient when the local agencies that can influence school cooperation and survey participation have a stake in the quality of the results.

4. DATA COLLECTION AND PROCESSING COSTS

NAEP and NELS share a common problem with respect to costs. Each study is already quite large by Department of Education standards for statistical research; they are two of the most expensive projects in the Center for Statistics. Yet, in important respects, the samples for the two studies are too small to accomodate several critical goals that have emerged from the growing public interest improving the nation's schools. Given the comparatively small size of the Education Department's total research budget, the problem of how these studies can be expanded to meet pressing new information needs appears, at times, to be unsolvable. This section examines two broad approaches to handling vexing cost problems: supplements to the studies by state education agencies, and design changes that offer increased efficiency in data collection and processing.

Costs of Obtaining State Level Data

Demands are frequently placed on both NAEP and NELS for data on specific subpopulations (e.g., handicapped students, bilingual students, students in specific types of vocational programs, from certain geographical regions, and so on), or at a level of precision they cannot now provide. Data for state level comparisons has become the most nagging shortcoming. These value of these comparisons for assessing the changes states are making in curriculum content and organization, graduation and course taking requirements, promotion policies, teacher certification, and many other areas will certainly increase over the next decade.

The usefulness of both NAEP and NELS would be enhanced if state level samples were designed into the Federally sponsored core study. But this would be a major shift in the objectives of the studies. At present, national population estimates are the primary goal of both projects. Estimates for subpopulations of interest (language minorities, handicapped students, school dropouts, etc.) are also sought on a national basis. As for any national sample survey, the great variability across states in the sizes of their (student) populations profoundly affects the sample design. Collecting data from 50 separate state-representative samples is an extremely inefficient way to obtain national statistical estimates. Under the current design, the state school and student samples for NAEP and NELS roughly reflect the very uneven distribution of students by state in the national population. With total student sample sizes of 25,000 to 30,000, under a proportional sample allocation, very few states have samples large enough to be considered representative. This sample design is both appropriate and efficient for the stated objectives. If the objectives shift toward placing higher value on state comparisons, the sample design should be altered to reflect those preferences.

The historical NAEP bias against state and district level comparisons, and NELS' lack of representative data for all but the few largest states deprive the studies of much of their potential impact on school improvement. For example, in the current plan for the NELS:88 Base Year survey, only eight states have as many as 30 public schools in the "core" sample (yielding about 700 sampled students), and only three states have more than 50 schools (with 1,200 or more sampled students). To

minimize survey costs, students within any state are not sampled directly, but are randomly selected from within samples of schools. Clustering student samples within schools reduces the relative efficiency of the estimates from the sample below what would be obtained from a simple (unclustered) random sample of the same size. The precision of national statistical estimates made with NELS data is generally about 15 percent lower than for more costly simple random samples of the same size. Efficiency losses for states are approximately comparable. While these losses are not large, they nevertheless reduce the "effective size" for the national and subnational samples.

There are few doubts about either the states' need for data, or the capability to conduct primary analyses at the state level. States with sufficiently large NELS core samples (especially New York, Illinois, Ohio, California, and Washington) have generally exploited HS&B data for the study of school dropouts, tracking and course taking patterns, effects of school size and structure on achievement, vocational and handicapped education programs, postsecondary school attendance, the impact of student employment on learning, and many other topics. Other states without enough sample points to permit the construction of separate analysis tapes may be consumers of NELS research reports and statistical publications, but lacking separate data on their own students and schools, they do not consider the paybacks sufficient to justify the costs of analyzing the datafiles themselves. This valuable resource for state policy analysis is thus likely to remain underutilized.

The majority of chief state school officers and state research directors have asserted that state representative NELS data would strengthen their capabilities. Although it was not always so, state-comparable assessment data is now widely regarded as highly desirable. The main difficulty facing these administrators is the widespread lack of state and local funds to collect and process the data.

Assuming that agreement between Federal and state officials on the objectives and procedures for learning assessment can eventually be negotiated, piggy-backing state level assessments onto NAEP offers the advantages of state-comparable data as well as cost containment for both state and Federal sponsors. The supplementation scheme offered by CS for NELS studies provide a reasonably similar framework for estimating costs. States sponsoring supplementary samples would avoid most of the large, up-front fixed cost component of establishing an assessment program, and would be responsible only for the costs of developing test or questionnaire items to be used only on their populations. Implementation costs for supplements are generally limited to the staff labor and other expenses (e.g., printing documents; travel, telephone, and postage; data preparation, conversion, and file construction) required to conduct additional survey and test sessions and process the data into analysis tapes. The costs for state supplementation of NELS:88 may provide a useful starting point for estimating the costs of state supplements to NAEP.

First, CS or contractor staff estimate the number of schools and

students expected in the Federally funded core sample within each state. With a total of 1,000 eighth grade schools in the NELS:88 core, and with schools selected with probabilities proportional to eighth grade enrollment, the estimated number of schools in each state ranged from 2 in the smallest states to about 100 in the largest (see Table 1). A fixed number of students (24) are to be selected for the core sample within each school, giving a range of 48 to about 2,400 students within states.

State requirements for sample sizes will depend upon their research goals, the expected magnitude of the program effects they may want to detect, and the statistical precision needed. To assist states in making choices, charts are prepared for each state showing the estimated sampling errors (adjusted for clustered sampling) associated with the school and student samples in each state. Rough estimates of the costs of supplementing the core sample with additional school administrations are also supplied in the initial correspondence.

Table 1. Estimated State Sample Sizes and Std Errors for Schools and Students
Eighth Grade Students in Public Schools Only

State	N of Schools	Sampled Schools	Total Students	% of Total	Std's/ Schl	SE 50 %	SE 90 %	95% Confidence
AK	228	2	7,973	0.2%	35	8.09	4.86	15.86
AL	553	14	57,877	1.8%	105	3.49	2.09	6.84
AR	402	9	36,551	1.1%	91	4.35	2.61	8.52
AZ	302	10	41,916	1.3%	139	4.15	2.49	8.14
CA*	1,608	81	326,014	10.2%	203	1.50	0.90	2.95
CO	329	11	44,414	1.4%	135	4.03	2.42	7.90
CT	222	9	37,844	1.2%	170	4.40	2.64	8.62
DC	35	2	6,155	0.2%	176	10.92	6.55	21.40
DE	31	2	7,733	0.2%	249	9.79	5.87	19.19
FL*	426	32	127,164	4.0%	299	2.42	1.45	4.73
GA	418	24	95,862	3.0%	229	2.78	1.67	5.45
HI	59	3	11,967	0.4%	203	7.85	4.71	15.39
IA	494	10	40,299	1.3%	82	4.11	2.46	8.05
ID	156	4	16,151	0.5%	104	6.60	3.96	12.93
IL*	1,478	36	142,398	4.4%	96	2.21	1.33	4.34
IN	463	20	80,800	2.5%	175	3.01	1.81	5.91
KS	451	8	31,375	1.0%	70	4.59	2.75	8.99
KY	525	13	52,417	1.6%	100	3.65	2.19	7.16
LA	468	16	62,334	1.9%	133	3.40	2.04	6.67
MA	400	18	72,011	2.2%	180	3.19	1.92	6.26
MD	225	13	53,407	1.7%	237	3.72	2.23	7.30
ME	244	4	17,338	0.5%	71	6.18	3.71	12.12
MI*	763	33	133,054	4.2%	174	2.35	1.41	4.60
MN	485	15	61,242	1.9%	126	3.42	2.05	6.71
MO	656	16	64,813	2.0%	99	3.28	1.97	6.44
MS	339	10	40,311	1.3%	119	4.21	2.52	8.25
MT	363	3	11,827	0.4%	33	6.51	3.91	12.76
NC	577	23	92,104	2.9%	160	2.82	1.69	5.52
ND	287	2	9,739	0.3%	34	7.26	4.36	14.23
NE	527	6	22,035	0.7%	42	5.08	3.05	9.95
NH	139	4	14,223	0.4%	102	7.03	4.22	13.77
NJ	698	23	93,382	2.9%	134	2.78	1.67	5.45
NM	155	5	20,846	0.7%	134	5.88	3.53	11.53
NV	85	3	12,523	0.4%	147	7.62	4.57	14.93
NY*	1,043	54	214,980	6.7%	206	1.85	1.11	3.63
OH*	890	37	149,483	4.7%	168	2.21	1.33	4.34
OK	671	13	50,197	1.6%	75	3.65	2.19	7.16
OR	372	9	36,578	1.1%	98	4.37	2.62	8.57
PA*	728	35	139,301	4.4%	191	2.30	1.38	4.51
RI	52	3	11,136	0.3%	214	8.15	4.89	15.97
SC	266	13	50,412	1.6%	190	3.82	2.29	7.49
SD	340	2	9,822	0.3%	29	6.86	4.12	13.45
TN	581	17	68,297	2.1%	118	3.23	1.94	6.33
TX*	1,466	62	249,433	7.8%	170	1.71	1.03	3.36
UT	133	7	28,202	0.9%	212	5.12	3.07	10.03
VA	341	20	80,835	2.5%	237	3.03	1.82	5.93
VT	134	2	9,057	0.3%	68	8.51	5.11	16.68
WA	406	15	58,692	1.8%	145	3.52	2.11	6.89
WI	535	15	61,389	1.9%	115	3.40	2.04	6.67
WV	285	7	28,849	0.9%	101	4.93	2.96	9.66
WY	110	2	7,890	0.2%	72	9.18	5.51	17.98
Total	22,944	800	3,199,965	100.0%	139	0.48	0.29	

Upon request, interested states will receive more detailed information on estimated sampling errors for individual subpopulations (e.g., racial or ethnic groups, sex categories, public vs. private school students), and for subgroup comparisons within their states. Table 2 displays the effect of increases in state sample sizes upon sampling errors for state population and subgroup estimates in a large northeastern state. Finally, a table of costs is provided showing the total costs for supplements of different sizes (Table 3).

Actual costs of state supplements to NAEP can only be determined by zero-based budgeting from clear specifications of the work scope. However, the data collection tasks for NAEP and NELS are basically similar, so estimated NELS:88 state supplement costs may provide a useable approximation of NAEP costs to states. Establishing and maintaining management and reporting procedures for each state supplement imposes a fixed cost of about $12,000 in each state. The cost of collecting and processing data in each school added beyond the core sample will be approximately $1,500 (about $1,000 for drawing the student sample, arranging and conducting survey activities, followup as necessary, and quality control supervision, and collection and processing of a school principal questionnaire; plus about $20 additional costs for collecting and processing data for each of the 24 students per school). Developing, collecting and processing an additional two pages of questionnaire items for students within a state averages about $15,000. (The average is based on an assumption of about 20 schools and 480 students in the core plus a state supplement of about equal size.)

Table 2. State Sample of Eighth Grade Schools and Students

No. of Schools	Schools In Sample	Total Enrolled	Sts / School	SE 50 %	95% Conf Int
1,146	60	200,000	174.52	1.81	3.55
1,146	65	200,000	174.52	1.74	3.41
1,146	70	200,000	174.52	1.67	3.28
1,146	75	200,000	174.52	1.61	3.16
1,146	80	200,000	174.52	1.56	3.06
1,146	85	200,000	174.52	1.51	2.96
1,146	90	200,000	174.52	1.47	2.87
1,146	95	200,000	174.52	1.42	2.79
1,146	100	200,000	174.52	1.39	2.72
1,146	105	200,000	174.52	1.35	2.65
1,146	110	200,000	174.52	1.32	2.58

Sampled Students	Whites	Blacks	Hispanics	SE 50% Whites	SE 50% Blacks	SE 50% Hispanics
1,620	1,053	324	243	2.01	3.17	3.49
1,755	1,141	351	263	1.93	3.05	3.35
1,890	1,229	378	284	1.86	2.94	3.22
2,025	1,316	405	304	1.79	2.83	3.11
2,160	1,404	432	324	1.74	2.74	3.01
2,295	1,492	459	344	1.68	2.66	2.92
2,430	1,580	486	365	1.63	2.58	2.84
2,565	1,667	513	385	1.59	2.51	2.76
2,700	1,755	540	405	1.54	2.45	2.69
2,835	1,843	567	425	1.51	2.39	2.63
2,970	1,931	594	446	1.47	2.33	2.56

SE Wh-Bl	SE Wh-Hsp	SE Bl-Hsp	95% Conf Wh-Bl	95% Conf Wh-Hsp	95% Conf Bl-Hs
3.76	4.03	4.71	7.36	7.89	9.24
3.61	3.87	4.53	7.07	7.58	8.87
3.47	3.72	4.36	6.81	7.30	8.55
3.35	3.59	4.21	6.57	7.04	8.25
3.25	3.48	4.08	6.36	6.82	7.99
3.15	3.37	3.95	6.17	6.61	7.75
3.06	3.28	3.84	5.99	6.42	7.53
2.97	3.19	3.74	5.83	6.25	7.32
2.90	3.10	3.64	5.68	6.08	7.13
2.82	3.03	3.55	5.53	5.93	6.96
2.76	2.96	3.47	5.40	5.79	6.80

The total cost of a state supplement of 65 schools with 1560 students and two additional pages of questionnaire items would total about $125,000 ($12,000 fixed cost, plus 65 * $1500 = $97,500 for supplementary schools, plus $15,000 for the questionnaire supplement = $124,500). Average costs

Table 3. Estimated Costs of State Level Supplements to NELS:88

Number of Added Schools	Without Questionnaire Supplement	With Questionnaire Supplement
5	$19,500	$34,500
10	$27,000	$42,000
15	$34,500	$49,500
20	$42,000	$57,000
25	$49,500	$64,500
30	$57,000	$72,000
35	$64,500	$79,500
40	$72,000	$87,000
45	$79,500	$94,500
50	$87,000	$102,000
55	$94,500	$109,500
60	$102,000	$117,000
65	$109,500	$124,500
70	$117,000	$132,000
75	$124,500	$139,500
80	$132,000	$147,000
85	$139,500	$154,500
90	$147,000	$162,000
95	$154,500	$169,500
100	$162,000	$177,000

per case (averaged across total fixed and variable costs) would be under $80, less than one-third the average per-case costs for the core sample (that must include all forms development and design costs in its total). For a supplement of 100 schools and 2,400 students, per case costs would

drop to-about $70 (fixed management and development costs would be a smaller proportion of total costs).

Costs for NAEP supplements might differ. Historically, much less data has been collected and processed for each NAEP student, but the data collection and processing procedures for NAEP are more expensive, and the average number of students per sampled school has generally been larger. Under the current design, NAEP document printing costs are more expensive than for NELS because of the complex requirements of BIB spiralling. Further, NELS standardized questionnaires permit the use of accurate, economical optical scanning technology for data conversion, whereas NAEP documents require computer assisted data entry. Data tapes may include more items for NELS, but have more complex structures for NAEP. Because of these offsetting features, actual supplementary costs for NAEP are not likely to be very much different from those for NELS, unless the work scope is defined very differently.

Some efforts to estimate supplementary costs to states have produced figures below $10 per student. This level of effort clearly does not cover either the range of activities described above, or adherence to all the steps required by the quality standards set for NAEP and other CS studies. The work scope could be defined quite differently for NAEP. For example, larger student samples might be assessed within core sample schools at costs of $10 to $20 per student. This approach avoids fixed costs gaining the cooperation of additional schools, drawing additional samples of students, conducting additional assessment sessions, and so on.

Increasing the cluster size (number of students per school) further would probably not be very harmful to the statistical efficiency of traditional NAEP analyses, which aggregate data across state boundaries. But the overall gain in assessment data useable by states would be relatively low under this approach. To support state comparisons and within state analyses, the NAEP design would have to move more in the direction of a standard rectangular data structure for students within a state, at least for background data if not for all assessment exercises. With this type of design, the statistical efficiency of adding more students in a core sample schools is far lower than that obtained from adding new schools and students to the Federal core. Cost estimates for NELS:88 are based upon adding schools (clusters) and are therefore probably more appropriate indicators of the level of effort required.

Improving Data Collection and Processing Efficiency

Throughout the histories of the NAEP and NELS programs funding shortages and the contract and grant procurement process have applied continuous pressure to increase the efficiency of each study. Under current research objectives, the discovery of new techniques or procedures that would produce for either study the same quantity and quality of data for significantly lower costs, or that would increase data output for the same cost, is unlikely. The following section describes some technologies that may greatly increase the capabilities of either study, and may increase the efficiency of specific tasks within the studies, but would require some increase in costs, at least initially. The only remaining

approaches for improving productivity may be in the consideration of merging aspects of sampling, forms design and data collection operations for two studies.

The scope of work for NAEP and for most waves of NELS surveys includes several identical elements (questionnaire and test development, sample design and selection, state-level coordination, gaining access to schools and students, conducting group survey sessions in school classrooms, data processing and file construction) that, from some perspectives, are redundant, duplicative, and that contribute to increased response burden. At the risk of reduncancy in this discussion, under the currently defined goals for NAEP and NELS, there is in fact very little duplication of effort between the studies. However, certain radical changes in the specifications for the studies could allow significant reductions in the need to conduct these activities separately. In effect, a single study would be created with a broader set of objectives preserving the essential elements of the separate projects. The number and costs of the redundant activities described above would be significantly reduced and the integration of studies sought by OERI would be greatly advanced. Although space does not permit full development of the implications for both studies, some thoughts on the benefits and problems of a merged design are outlined below.

1. Establish grade-level sampling for both surveys. Design and select single, coordinated samples of schools, students, teachers and parents from whom combined NAEP and NELS data would be collected.

Samples at the 4th, 8th, and 12th grade levels would be selected as assessment cohorts and as Base Year grade cohorts for three parallel longitudinal surveys. To serve current longitudinal studies goals for national estimates, the number of schools should be about 1,000 per grade or 3,000 in all. (In practice, some schools might be selected for more than one grade cohort, but the probability of this is small and could be eliminated, if desired, by a design restriction.) NELS samples would normally include 1,000 schools for a grade cohort, and NAEP about 1,700 different schools for three age-grade cohorts. The integrated school sample in the first study wave (Time-1) would therefore be only slightly larger than the total number of schools that would be selected under current plans for the separate studies whenever a NELS school-based survey occurs in an assessment year (i.e., 1988, 1990, and 1992 in the forseeable future). Sensitivity studies could be performed to indicate the statistical losses to NELS objectives if the sample sizes were cut to 900 schools per grade level and the within-school cluster size increased somewhat.

Costs of data collection from additional schools might be offset several ways. First, construction of the sampling frames and selection of the school sample would be much more efficient if done for all grades by the same firm. Second, under this design, it would desirable to introduce public school districts or groups of districts (and other administrative units for private schools) as stages of selection (primary or secondary sampling units). Restricting the study to 1,000 or fewer total locations would greatly reduce the costs of gaining cooperation of districts and

schools as well as the travel costs for survey and test administrators. It would also avoid increasing the complexity and costs for longitudinal followups of students moving through the elementary and secondary grades by maximizing the overlap in location of followup surveys.

Finally, and perhaps most important, by collecting more data from the same units, the total number of student surveyed in a typical NAEP-NELS overlap year (over 150,000 in 1979-80 and over 114,000 expected in 1987-88) could be greatly reduced. If the number of students per grade is no more than 24,000 as for NELS:88, the total student sample would be 72,000 for the first year of the survey. This sample size would permit the inclusion in the survey of one parent per sampled student, from whom essential background, socio-economic status, and home environment data can be obtained, without exceeding the total number of respondents surveyed by jointly by NAEP and HS&B in 1979-80. However, since the unit costs of surveying students and parents is not the same, it may not be possible to survey parents of all three cohorts during the same year. It is important to note that, once collected at Time-1, data from parents in a merged study design would be available for analysis of multiple waves of student data and need not be collected in all subsequent waves.

Many challenging complications arise for the next assessment or longitudinal followup two years later (Time-2). Most members of the NELS grade cohorts would have progressed to the 6th and 10th grades, or would be two years beyond high school; no previously selected students would be in standard NAEP grades. Fresh samples could be drawn from the 4th, 8th

and 12th grades, but in the absence of other reductions, this would double the number of students being surveyed at Time 2, considerably raising costs in comparison to the current separate designs. One or more of the longitudinal studies cohorts might be dropped, or all three reduced in size, but this would defeat many of the purposes of the design change.

Under a radical but less costly alternative, NAEP assessments at Time-2 could be carried out on the longitudinal cohort members in the 6th and 10th grades, on dropouts and on those who left school, adding a new dimension to the scope of NAEP products and reports. Two years later (Time-3), most of the original NELS 4th and 8th grade cohorts will have reached the 8th and 12th grades and may be reassessed. However, these Time-3 samples will represent only 8th and 12 graders who were in 4th and 8th grades at Time-1, not the full populations of 8th and 12th graders of that year (which will include newly arrived immigrants as well as students not promoted on schedule. A similar problem is already being faced by NELS:88. Procedures are being designed to augment the 8th grade cohort of 1988 with a supplementary sample of 10th graders in 1990 in order to insure comparability with the HS&B 10th grade cohort of 1980.

Another serious complication is population mobility and other causes of school change by students, which will reduce the number of originally sampled students who may be assessed and resurveyed at the same school in subsequent years. In the absence of empirical data, these problems are easy to exaggerate. The problem of school change between the 10th and 12th grades was relatively minor for the first followup of HS&B, but this

time span was short and covered a period where school-change is uncommon. The problem for NELS:88 will be much greater, and the solutions developed to handle it more relevant to the problems of a merged study.

Finally, at Time-3 it will be necessary to draw a new sample of 1,000 4th grade districts and schools to restart the cycle. These selections should be made from different geographical areas in order to rotate the survey burden across all eligible units. The total number of schools in the sample would thus remain at 3,000, but the number of geographical areas would increase somewhat, depending upon the specific sites selected.

The total number of students surveyed at Time-3 would also increase. The size of the increase would depend upon the extent of the reduction in sample size for the eldest cohort (12th graders at Time-1) after leaving high school (Time-2). The generally higher per-case costs of out-of-school survey administration has resulting in reduction of HS&B cohorts to 12,000 to 15,000 (less than half their initial size) after high school graduation.

2. Collect all essential NAEP and NELS data from the same schools, students and other respondents. Integrate survey questionnaires and tests to a much higher degree than under the current design.

Questionnaires of NELS length and content coverage would supply virtually all items needed for traditional NAEP analyses and would greatly expand the number and range of possible analyses of the assessment data.

Integrating test construction (and administration) will be more complicated because of the numerous differences in the current objectives of the two test programs. Vast arrays of NAEP items are divided among respondents so that only small subsamples take the same tests, and no respondent encounters more than a small proportion of items in the subject pools. NELS collects a smaller amount of identical test information from every sampled student. NAEP reassessments rarely occur at intervals of less than four years, while NELS must test the same subjects every two years. Historically NAEP assessments have been limited to two or three subjects or skills per year, whereas NELS has included at least four subject areas in its test batteries. NAEP item pools must be retained for comparability with past assessments. NELS also has a cross-cohort component that should not be ignored. The NAEP item pools may prove adequate as sources for NELS tests. Although in principal it should not be necessary, additional items will probably be needed to meet NELS' objectives of measuring cognitive growth at the individual level.

Under the alternative survey plan described above, item pools would have to be created for 6th and 10th grade assessments. Although the HS&B study design included tests designed for 10th graders, the tests included a relatively small number of items, few of which would be suitable for assessment purposes under this design.

To prevent undesirable erosion of the capabilities of the two approaches, two types of tests and two testing sessions might be

necessary. Arranging two testing sessions runs counter to the purpose of reducing costs, but provides opportunities for increasing the research potential of a merged study. If school samples were drawn early enough and their participation arranged during late summer preceding the school year for the study, a fall testing session might be arranged. (Until recently, the assessment schedule for 13 year olds was from October through December of the assessment year.) This session would replicate existing or modified NAEP procedures, and could probably expand the number of test items per student because little or no background data would need to be collected in this session. In the spring, a second session would be held, modelled after the current NELS approach. A 45 minute questionnaire would collect all needed background data. A second one hour test battery (identical items for all students), optimized for the measurement of growth across school grades, would be administered. The two test documents would include a subset of common items in specified subject areas designed to measure gross levels of growth during the school year.

If the tests could be fully integrated, the second test would consist entirely of items included in the pools for the first (fall) test, permitting the study of growth and learning during the school year. In practice, this is not likely to be possible if both NAEP and NELS traditional testing goals are retained. However, the use of Item Response Theory results to construct similar individual level composite achievement scores from both the complex assessment data matrix and the item scores measured in the spring test offers some hope of estimating growth within the school year. In addition, if the spring tests for all three grades

are constructed to measure growth for longitudinal cohort members, they ,may also be suitable for synthetic cohort analyses after each wave of data collection was completed.

3. Process assessment and longitudinal studies data consistently, combining both types in a relational database permitting the creation of tailored analysis files and public use data tapes on demand.

Aside from the mandated performance "report cards" and the new policy analyses introduced into the current grant cycle, very little use is made of NAEP data to determine how educational processes might be improved at any level. The main reason for the low level of utilization is that so little additional data is available to support data analysis at the level of the individual, school, or even the type of school at which assessment data are obtained. Facing rapidly increasing need for data to address policy issues, the Education Department must soon answer the question of how long the luxury of an expensive, single-purpose educational indicators program can be afforded.

In a merged study, far more individual and lower level aggregate data will be available to exploit assessment results. Various types of assessment scores and composite variables could be generated from the first (fall) testing phase of a combined study and added to the typical data record produced for NELS studies. Although the combined costs of processing both NAEP and NELS data elements in a merged study will be cumulative and will offer few cost efficiencies, the total costs of merged data file production should be somewhat lower if the tasks are done concurrently, compared to the cost of attempting to create appropriately

merged datasets under current conditions (assuming the population definitions were consistent).

A relational database provides the most obvious means for linking data elements available for different levels or aggregates. Just as the current BIB design permits correlations to be calculated between many assessment exercise that are paired in blocks, relationships among assessment scores and a wide range of background, educational activity, and environmental variables can be calculated as well. Case identification numbers and assessment block assignments will serve as the primary indices for linking the test and questionnaire data. A crucial difference from current practice will be that all forms of background, environment, and process data will be available for all students, regardless of assessment block assignment. The great increase in the amount of common data for sample members may improve procedures for imputing particular assessment exercise scores for cases not assigned to a particular block, helping to simulate a rectangular assessment data matrix.

The publication of thorough documentation and data file users manuals of the kind typical of NELS studies would then cover assessment data as well as longitudinal data. The elimination of completely separate data systems, manuals, documentation and related publications for NAEP and NELS will eliminate some duplication of effort and reduce costs slightly. Dissemination efforts for the merged dataset would be more focused, and would reach a wider audience of prospective analysts.

5. OPPORTUNITIES PROVIDED BY NEW TECHNOLOGY

The increasing capability and decreasing cost of microcomputers are strongly influencing how survey research is done. They are now used routinely for tasks that formerly required large mainframe computers, including: constructing sample frames and selecting large, complex samples; developing and revising survey questionnaires, cognitive tests and related documents; computer-assisted telephone interviewing (CATI) and personal interviewing (CAPI); project scheduling, receipt control, survey management, and cost monitoring; computer-assisted data entry (CADE); datafile construction and documentation, and database management. Several of these functions are already in use on NAEP and NELS studies, and the remainder have obvious applicability.

Two additional applications have special relevance for both studies — in their present form as well as under any design modification — computer assisted testing (CAT), and computer assisted interviewing (CAI), by which I mean the self-administration of survey questionnaires. In computerized administration, keyboards, microchips, memory circuits, and display screens would replace the soft-lead pencils, and paper booklets now used by NAEP and NELS. In principal, completing a self-administered questionnaire at a computer station should be easier, more enjoyable, less burdensome and less susceptible to error by students than the traditional method of administration. Computer assisted testing is being used effectively in many personnel, research, and evaluation programs to improve efficiency in obtaining and using ability, aptitude, and

achievement scores from people of all ages, educational levels and occupations. Microcomputers could assist the NAEP and NELS testing programs in many of the same ways, including improving test security and speeding the calculation and reporting of individual questionnaire and test results.

Perhaps most important, however, is the enormous power that CAT techniques offer test designers to control the selection and presentation of test exercises to sampled students. Control may be exercised over the distribution of subsets of items to random subsamples as required by NAEP (computer assigned testing). Or it may be used to choose a set of items best suited to the measured ability of the examinee, in order to improve the accuracy and precision of individual assessments, as desirable for NELS:88 (computer adaptive testing). Additional control of this kind has the potential for increasing the efficiency of testing, providing more information for the same cost.

Computer Assisted Testing

NAEP's purpose is to obtain estimates of individual item scores (or groups of items), and some limited data on the correlations among scores on individual items. NAEP seeks breadth in its measurement of attainment levels for specific subject areas in the population. Except in a very limited sense, the current NAEP design does not aim to produce accurate assessments of individual students.

NELS:88 seeks to discover the factors that influence the growth and decay of educational attainment in students. Except in a very limited sense, NELS:88 cannot accurately assess the attainment of the national populations of 8th, 10th, and 12th graders on the broad fields mathematics, science, language arts and social studies. The continued use of paper and pencil test administration not only limits the ability of the two studies to meet their primary goals, it also ensures that the limitations on expanding those goals will remain in place indefinitely.

Even if, as I have suggested, individual response burden for NAEP is relatively light and might be increased in the interest of obtaining needed data, the room for expansion is certainly limited. Individual respondent burden for NELS studies is now about twice that for NAEP, and is probably close to the upper limit. The amount of time allotted for testing students in the two studies is quite short (about 45 minutes for NAEP; up to 75 minutes planned for NELS:88). Within these time spans, numerous objectives must be satisfied, including:

1. Students must be tested in multiple subject areas; in NAEP they must be tested on a large number of subtopics and skills within subject areas;

2. In each test administration, measurement accuracy in for individual items, groups of items, individual respondents, and groups should be as high as possible (effects of guessing and ceiling and floor effects should be minimized);

3. Measurements should permit analyses of performance differences between groups and changes in time-series measures of aggregate scores (NAEP), and changes in individual performance over several grades (NELS);

4. Tests for both studies should include items permitting bridges between NAEP and NELS as well as across NELS studies;

5. Both within and across the two studies, test scores must be scalable on the same metric.

To meet the first objective within the time constraints, subject area tests must be quite short (15 to 20 minutes) for NELS studies, and individual assessment exercises must be distributed in small subsets across sampled students in NAEP. Item Response Theory (IRT) scaling has contributed to the resolution of the fourth and fifth objectives. Meeting the second and third objectives within the time constraint presents the greatest problems for two studies; obtaining measurement accuracy at both the item level and the individual level (for status and change measures) becomes exceptionally difficult when the range of ability in the population is large and the number of items included in the test for each individual is small.

Thus far, NELS studies have tried to meet these objectives with standard linear tests that present the same items to all students in a cohort. While this maximizes the amount of information available for each test item, it severely limits the number of items per subject area and hence the amount of information gained about the achievement of students of all ability levels. In NAEP, matrix sampling and now BIB spiralling are used to cope with the large number of exercises that must be included in each subject or skill assessment. Although students take tests containing different combinations of items, the different test forms are all used within the same grade level, they all contain similar mixes of test items at each difficulty level, and they are distributed to students without regard for the students' ability levels.

Even without other constraints, conventional testing procedures that

present the same set of items (or, as in NAEP, an equally difficult set of items) to all students are usually less accurate in measuring the achievement of students at either extreme of ability. For students who are much above or below average, all the items presented may be much too easy or too difficult. As a result, unless the difficulty range of the test items is very great, a fixed- or equivalent-format test will not detect much variability in performance for students of the highest and lowest ability. This problem is of relatively little concern in NAEP, where only cross-sectional or time-series population estimates, not individual data, are needed. It is greatly compounded in NELS:88, with its purpose of measuring individual growth as students move from the 8th through the 10th to the 12th grade. For NELS:88, rather than assign the same test items to all students, it would be better to adapt the range of difficulty of the items presented to the ability level of each student.

To measure longitudinal change, the current proposed solution for NELS:88 is to construct three tests batteries (one each for grade 8, 10 and 12), all based on item pools large enough to cover the content and difficulty ranges for all three grade levels. Each grade-level test would contain subsets of items that overlap with the tests for adjacent grade levels, reinforcing construct stability and establishing a continuous range of item difficulty across the three tests. The design is thus adaptive at grade level. However, at each level all sampled students will take the same test, regardless of ability. The overlap of items for adjacent grade levels should decrease the problem of ceiling and floor effects at each grade, but expanding the overall difficulty range of each

grade-level test will tend to reduce the precision in the measurement of individual achievement scores. Significantly improving the measurement accuracy for students within each grade would require a means of presenting to each student a range of exercises suited to his or her ability.

In order to cover the very large number of topics, NAEP must design, produce and control the assignment of a large number of different test forms (63 separate booklets in the 1983-84 assessment) according to a complex allocation scheme. For all its complexity, the number of booklets (blocks of items) still had to be limited to a number that could be printed and distributed by assessment teams in school classrooms. Computerized administration not only provides a better means for physically presenting the blocks of exercises to students, it also permits much greater flexibility in the creation of blocks than paper booklets could support — for example, by allowing random variation of the order in which the items in a given block is presented.

Although BIB designs and other experiment-related assignment procedures accomplish current NAEP goals for item distribution, they reflect a specific mix of statistical analysis and reporting objectives. Other mixes, for example, one that would put a higher emphasis on collecting efficient individual level data, might require the adaptive assignment of later items in the test sequence to be conditional on examinees responses to the earliest items presented, as is desirable for NELS. The fact that adaptive tests can gather more information about

student achievement in less time than standard linear tests might make the option of collecting individual level data more attractive because it imposes smaller opportunity costs.

NELS:88 faces the entirely different problem of measuring achievement scores for students of widely varying ability with sufficient accuracy to assess change in achievement across two-year and four-year intervals. Success depends upon the accurate measurement of individual attainment at each survey wave. With the aid of Item Response Theory (IRT) scaling, computerized test administration could select the items presented to a student from a large pool of items, using information about the students' success in answering all prior items. The CAT software thus adapts the composition of the test to the ability level of each student, raising both the precision and the accuracy of the measurement. The techniques for constructing and analyzing fully adaptive tests have been worked out by several firms, and are reportedly in use in Department of Defense testing programs and in private industry.

Beyond the efficiency gains of control over item selection and presentation, computer assisted testing also permits the use of other test item formats besides multiple choice items now used for most NAEP and NELS test batteries. The need to expand data capture methods beyond scoring multiple choice items may be especially important for developing and administering tests of higher order thinking skills for use in the future.

Items of this kind may require entry of open-ended responses, or even symbol manipulation using the graphics capabilities of the console display screen.

Computer Assisted Interviewing

Computer assisted telephone interviewing has been widely used in Federal survey contracts for over two decades. Computer assisted personal interviewing, using portable laptop computers carried by interviewers into respondents' homes or other settings, has been specified for a number of large scale Federally funded surveys now getting underway. The next logical step is the application of the same technology to permit sample members to complete computer administered questionnaires by themselves, group settings, with only minimal supervision by interviewers or other field data collection staff. Coupled with computerized testing described above, computer assisted interviewing has the potential to introduce the first major jump in the efficiency of survey and test data processing since the emergence of mark sense optical scanning as a reliable, economical data conversion technique.

The advantages of computer assisted interviewing in studies like NAEP or NELS include, but are not limited to:

1. Ease of Administration.

Respondents of almost any age and ability level can be guided through a sequence of conditional branches in a complicated questionnaire,

eliminating response errors (e.g., failing to follow skip patterns; providing more than one answer when only one is desired, etc.), missing data caused by skipped pages, and respondent frustration related to misunderstood instructions; experimentation with item wording and format can be done without costly forms production, distribution, and processing problems;

2. Data Quality.

Out of range responses and consistency problems may be eliminated;

3. Instant Help to Respondents.

Interviewing software can provide, at the stroke of a key, immediate, context-sensitive help to respondents who are confused about the meaning of a question or the available response categories;

4. Instant Data Processing.

Responses are processed and cleaned as they are entered, eliminating printing and distribution costs, as well as the time delays associated with data entry or optical scanning . As soon as the interview is completed, the data are ready to be merged with other records in analysis files. Data from each respondent can be processed on a flow basis, rather than in batch mode;

5. Input of Text, Numeric, or Unconventional Responses.

Respondents with minimum typing skills can enter "open-ended" (i.e., verbatim) responses directly via the keyboard, dollar amounts and dates

can also be entered without the use of cumbersome grids of ovals for scanning equipment or error prone data entry;

6. Facilitating Data Collection from Handicapped Students.

The use of computer controlled audio devices (tape recorders, voice synthesizers, or other equipment) and "braille" keypads could permit data collection from visually handicapped students. Other input devices might be used by students with physical handicaps that preclude paper and pencil menthods. Use of single computer controlled audio devices could also restore the pacing of testing sessions for all students lost under the current BIB design.

7. Handling Multiple Questionnaire Forms.

Circumstances may require the assignment of more questionnaire items than be presented to one respondent, as occurs in NAEP assessment batteries. While this practice has a number of serious disadvantages for analysis in longitudinal designs, it is easier and less costly to implement effectively by computer than by paper and pencil forms.

8. Administrative Control.

Since record keeping is integrated into the interviewing software, receipt control and survey management are made more accurate and cost effective.

Problems of Implementation

No purpose would be served by suggesting that the implementation of computer assisted interviewing and testing is a move of a short distance. Although most of the initial difficulties of the programming for required software have been overcome, a great deal of development work remains. Unquestionably, the most difficult problems will be

1) assessing whether the considerable gains in policy research capability provided by the computerized approach are sufficiently valuable to justify the additional cost; and

2) the development of workable procedures for bringing computers and students (and other respondents) together for data collection.

On the matter of costs and benefits, while the costs of hardware and remaining software development will be substantial at the beginning, evaluators should keep in mind the longer-term facets of this approach. Government funding of the CAT and CAI software will put these products in the public domain, making them available to any contractor possessing all other relevant capabilities for this type of work. Computer hardware purchased by the government will not be a repeating expense for each survey, but may be amortized over several assessments or longitudinal survey waves. Once the initial investment is made, however, it is virtually certain that additional uses for the same equipment will arise, spreading costs across more and more studies. The ratio of costs to benefits may be especially low for a study designed to merge NAEP and NELS goals through data collection from the same samples of students and schools. While this is not sufficient reason for favoring a merged study design, if that approach is favored for the purpose of advancing the research objectives described above, the use of CAT and CAI technology

would be exceptionally attractive.

To risk stating the obvious, the first use of computer assisted interviewing and testing will face all of the typically fierce difficulties of establishing a new system of data collection and processing. Under the pressure of time and financial limitations, not every problem will be forseen, and not all will be solved. If and when the decision to adopt this technology is made, it should be implemented in cautious feasibility and pilot studies before a commitment is made to a large scale application. This might be handled through separate evaluation contracts or grants, apart from NAEP or NELS studies proper, or in field test phases of the two projects, or by both means.

In addition to the functioning of all aspects of the CAT and CAI software, the most demanding issues will be how students and computers are most efficiently brought together. Under paper and pencil approaches, the amount of materials and hardware involved is minimal. NAEP and NELS have traditionally used group administrations in schools where students are enrolled. Although computers could be introduced into this approach, other arrangements are possible. For example, a central location in a population center could be established to handle students from many sampled schools in the area. Such a procedure might also have salutory effects upon sample design (lowering cluster sizes) as well as data quality, and might reduce the disruption of the instructional schedules of individual schools. Other possiblities, such as setting up a small number of fully programmed machines to remain in each sampled school over

extended time periods, offers other advantages; for example, the availability of computers would facilitate the capture of principal and teacher data, and other school records, as well as student questionnaire and test data.

In my view, while the costs of adopting computer technology to education research may be high compared to the overall level of investment in social research, the advantages conferred by the increase in data collection power will prove to be vastly higher. The Office of Educational Research and Improvement and the Center for Statistics have opportunities in both NAEP and NELS to chart a course in this direction, and a responsibility to the public to establish that it is worthwhile. Inherent in this choice is a "catch-22:" because the paypack period for adopting computer technology will extend over several years, the initial expenditure of funds will have to be made despite the risk that it may not work out in the long run. But this is not very different from similar commitments made by many other Federal research agencies. An investment in technology at this level must be regarded as something akin to the space shuttle program. Since the economic advantages can only be realized through repeated use of the hardware, the decision must be accompanied by a long term utilization plan to achieve the desired balance. Acknowledging the lesson of recent history, sufficient trials and development effort must be expended in pilot testing to gain the necessary expertise and to avoid disaster.

The analogy to the space program may be apt in another sense as well. The research and development costs expended to solve specific problems of NASA programs frequently resulted in spillover benefits to other areas of research less well endowed with R&D funds. The development of computer assisted techniques for studies like NAEP and NELS would almost certainly promote efficiency gains for many other school-based studies now operated or being planned by the Center for Statistics. While the focused effort of developing these capabilities should thus be confined to a single project or set of closely related projects, the benefits will be shared by many others that do not bear development costs.

Finally, just as the space program has evolved over its history into a partnership between U.S. Government interests and other actors (private firms, foreign governments), there is considerable evidence of interest by other actors (primarily state education agencies, but also including private companies) in the development of computer assisted survey and assessment technologies. As for all other aspects of survey and assessment practice, it would be highly desirable to establish common standards at this early stage in the consideration of computer applications. It would be especially desirable to form a standing Survey and Assessment Technology Committee to consider active collaboration and cost sharing on research and development efforts.

SUMMARY AND CONCLUSIONS

The need to integrate the various data collection and reporting activities of the Department of Education is well recognized by the current leadership, and attempts are being mounted to improve the linkages between the major studies it sponsors. Yet, at present, because of design dissimilarities, there is virtually no meaningful integration of the NAEP and NELS programs. In the preceding sections, I have tried to illustrate several of the most costly, wasteful, and frustrating aspects of this situation, and to suggest some possible means for correcting it.

Solutions to the problem, however, are not lying around in plain sight. In most instances, integrating the studies will require significant alterations to bring one of the studies into line with the other. Changes of this type are themselves disruptive and can be costly at the outset. By now, both NAEP and NELS have established relatively long research traditions; radical change will threaten the capacity for trend analyses important to both. Perceiving the benefits they might offer requires a longer-term and less "proprietary" perspective than the designers and operators of each study are likely to evince.

Some few changes should be made to both studies -- less to improve their integration than to upgrade the quality of the research data each collects. Perhaps most important in this regard are changes in the NAEP sample design regarding age vs. grade level sampling, and the introduction of computerized testing methods to improve the efficiency of measuring

various kinds of differences in what students have learned. However, care should be taken in the application of technological "fixes" to the research integration problem. The introduction of a change so fundamental as computerized administration to only one of the two programs would certainly aggravate existing differences between the two.

Most of the changes suggested to enhance integration of NAEP and NELS come with substantial price tags. While in the view of many, the best solution for the general public would be for the Federal Department of Education to reorient its approach to survey and assessment data collection and enlarge its samples to provide for state and large district representativeness, this perennial cry for expanding the research resources may never succeed in gaining the necessary Congressional appropriations. However, even in the absence of full Federal sponsorship of research at these levels, there remains an opportunity for the Federal department to exercise initiative and leadership to form consortia with states and localities, and to adjust its own data collection objectives to be more compatible with state and local needs. This will require greater emphasis on the collection of individual level NAEP data than ever considered possible or desirable in the past.

In an important sense, all of the possible remedies discussed are minor adjustments compared to the radical alternative of combining the research objectives of the two studies into a genuinely integrated research program. From the scientific, methodological, and administrative points of view, there is no reason why NAEP and NELS could not be merged

into a single system of surveys, combining both cross-sectional and frequently re-started, longitudinal studies, using most of the NELS design features and questionnaires and adapting NAEP test formats and analytical theory. However, the recent history of attempts to establish integrated elementary and secondary education data systems suggests that proposals for revolutionary change will surely encounter many difficult political (as well as technical and methodological) obstacles. Most of the problems arise from the large size and consequent costs of creating a fully integrated program, the existence of strong, organized constituencies committed to the design and products of the existing separate studies, and real concern over the risk that a radical new departure might fail to serve critical information needs that the current studies handle reasonably well.

I have tried to keep in mind the most fundamental differences between the purposes of the two studies. NAEP is primarily a learning assessment program, whose major dependent variables are large numbers of separate test scores that may be analyzed and explained by other independent measures, but whose main purpose is to be conveyed in descriptive reports to many constituencies. NELS is, and should remain, a broad study of the educational process, its context, and its outcomes, whose purposes are as much analytical as descriptive, and within which student test scores are sometimes dependent variables, but even more often are independent (classification) variables used in the analysis of other, later behaviors. It is not clear to me that these purposes must be pursued by separately designed and executed studies, or that the two sets of objectives must be

fully merged to succeed in meeting the Department's goals for research integration. I believe, however, that the time is right for careful consideration of the opportunities for major changes in both studies that may enhance the ability of the Department of Education to serve as valued collaborator in the process of improving elementary and secondary education at the state and local level.